

Weekly Report

Pingping Shang

2013.11.11~2013.11.17

本周工作

1. 丁师兄报告的那篇论文是利用信息熵选取变量或者感兴趣区域，认为信息量大的区域也是我们需要关注的变量或区域。信息论中有提到，变量的不确定性越大，熵也就越大，把它搞清楚所需要的信息量也越大。

考虑到我们的用户标签和用户行为，用户属性很多，有些属性对人们的购买行为没有区分作用，比如星座，有些属性对购买行为区分很强，比如性别。这可以类比为那篇论文中的变量选择：

- 1) 选取信息量大的属性，即熵比较大的属性；
- 2) 某一属性下，选取熵大的 block。

我将展示人群的信息熵都计算出来：

gender		
entropy: 58.08		
block0: 30.04		
block1: 28.04		
age		starlevel
entropy: 136.34		entropy: 152.48
block0: 29.31		block0: 14.97
block1: 24.31		block1: 30.51
block2: 26.00		block2: 28.28
block3: 27.41		block3: 25.33
block4: 29.31		block4: 26.29
		block5: 27.10
		decade
		entropy: 104.00
		block0: 0.0
constellation		block1: 27.20
entropy: 297.18		block2: 26.18
block0: 24.29		block3: 23.97
block1: 24.75		block4: 26.64
block2: 24.03		
block3: 24.14		
block4: 24.91		
block5: 24.17		
block6: 25.58		
block7: 22.1		
block8: 26.01		
block9: 25.74		
block10: 26.61		
block11: 24.54		
		buygrade
		entropy: 161.46
		block0: 30.47
		block1: 25.45
		block2: 28.23
		block3: 25.59
		block4: 23.75
		block5: 27.96
		cameraman
		entropy: 51.73
		block0: 23.88
		block1: 27.85

说明--block0: 24.29 冒号后面的数值表示该 block 的熵，计算公式为：

$$H(x)=E[l(x)]=E[-\log(2,1/p(x))]=-\sum p(x_i)\log(2,p(x_i)) \quad (i=1,2,...n)$$

这里的变量 x 为某 block（即某个人群）的商品购买人数，购买具体某类商品的人数为一个变量值。

观察计算结果，gender 的熵值较大，符合我们直观想法（包含信息量大），age 和 constellation 的熵值没有很大区别，跟我们的预期不符。所以，是不是不能把对具体商品的购买人数当作变量值？

下周工作

探讨能否从信息熵方面入手区域选择，标签选择。